IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - ✓ Gather data using SpaceX REST API and web scraping techniques.
  - ✓ Manage data to create a success/fail outcome variable.
  - ✓ Investigate data using data visualization techniques, taking into account variables such as payload, launch site, flight number, and yearly trends.
  - ✓ Study the data using SQL and compute statistics like total payload, payload range for successful launches, and the total number of successful and failed outcomes.
  - ✓ Examine launch site success rates and their proximity to geographical markers.
  - ✓ Depict the launch sites with the most success and successful payload ranges through visualization.
  - ✓ Develop models using logistic regression, support vector machines (SVM), decision tree, and K-nearest neighbor (KNN) to predict landing outcomes.

Introduction

SpaceX promotes Falcon 9 rocket launches on its website, presenting a price tag of 62 million dollars, which is significantly lower than the cost of competing providers, whose offerings can easily exceed 165 million dollars per launch. A key factor contributing to SpaceX's cost-saving advantage is the ability to recycle the first stage of the rocket. As such, an accurate determination of whether the first stage will land is paramount in evaluating the overall cost of a SpaceX launch. This knowledge can prove invaluable to potential competitors in the bidding process for rocket launches. Within this project, we attempt to answer the following questions:

- How different factors impact the success of first-stage landing:
  - *Payload mass*
  - *Launch site*
  - *Number of flights*
  - *Orbits*
- The trend of the success rate of first-stage landings over time.
- The identification of the best predictive model for determining the success or failure of first-stage landing, with a focus on binary classification.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Rest API and Web scraping
- Perform data wrangling
  - Data Normalization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Binary classification
  - Confusion matrix

Data Collection

1. Collect data from the SpaceX REST API, specifically the api.spacexdata.com/v4/launches/past endpoint, using a GET request with the requests library.
2. Normalize the obtained JSON data into a flat table using the json_normalize function.
3. Web scrape related Wiki pages using the Python BeautifulSoup package to obtain Falcon 9 launch records.
4. Parse the web scraped data and convert it into a Pandas data frame for further analysis.
5. Filter out Falcon 1 launches from the data to focus on Falcon 9 launches.
6. Deal with null values in the PayloadMass column by calculating the mean and replacing the null values with it.
7. Leave the column LandingPad with NULL values as is, as it indicates the non-use of a landing pad and will be dealt with later using one hot encoding.

Data Collection

1. Collect data from the SpaceX REST API, specifically the api.spacexdata.com/v4/launches/past endpoint, using a GET request with the requests library.
2. Normalize the obtained JSON data into a flat table using the json_normalize function.
3. Web scrape related Wiki pages using the Python BeautifulSoup package to obtain Falcon 9 launch records.
4. Parse the web scraped data and convert it into a Pandas data frame for further analysis.
5. Filter out Falcon 1 launches from the data to focus on Falcon 9 launches.
6. Deal with null values in the PayloadMass column by calculating the mean and replacing the null values with it.
7. Leave the column LandingPad with NULL values as is, as it indicates the non-use of a landing pad and will be dealt with later using one hot encoding.

Data Wrangling

1. Review the attributes of the data, including Flight Number, Date, Booster version, Payload mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing pad, Block, Reused count, Serial, and Longitude and latitude of launch.
2. Identify the launch sites in the LaunchSite column, which include Vandenberg AFB, Space Launch Kennedy Space Center, CCAFS, and SLC 40.
3. Determine the different payload orbits in the Orbit column, such as LEO and GTO.
4. Convert the landing outcome in the Outcome column to classes 0 and 1, where 0 represents a bad outcome (the booster did not land) and 1 represents a good outcome (the booster did land).
5. Assign the classification variable Y to represent the outcome of each launch.

EDA with Data Visualization

Visualization Methods:

- ✓ **Scatter plots** can be used to visualize the relationship between variables, and if a relationship exists, it could be valuable for machine learning purposes.
- ✓ **Bar charts** are useful for comparing discrete categories and showing the relationship between each category and a measured value.

EDA with SQL

1. Conduct Exploratory Data Analysis (EDA) using a database.
2. Determine if the data can be used to predict if the Falcon 9's first stage will land.
3. Identify attributes that can be used to determine if the first stage can be reused.
4. Use machine learning to predict if the first stage can land successfully by incorporating the identified features.
5. Incorporate success rate and launch site information as features.
6. Combine multiple features for obtaining more information.
7. Convert categorical variables using one hot encoding.
8. Prepare the data for a machine learning model that will predict if the first stage will successfully land.

Build an Interactive Map with Folium

1. Use interactive visual analytics to explore data interactively.
2. Build an interactive map with Folium to analyze launch site geo and proximities.
3. Use Plotly Dash to build a dashboard with input components like dropdown lists and range sliders.
4. Use the dashboard to find insights from the SpaceX dataset more easily.

Build a Dashboard with Plotly Dash

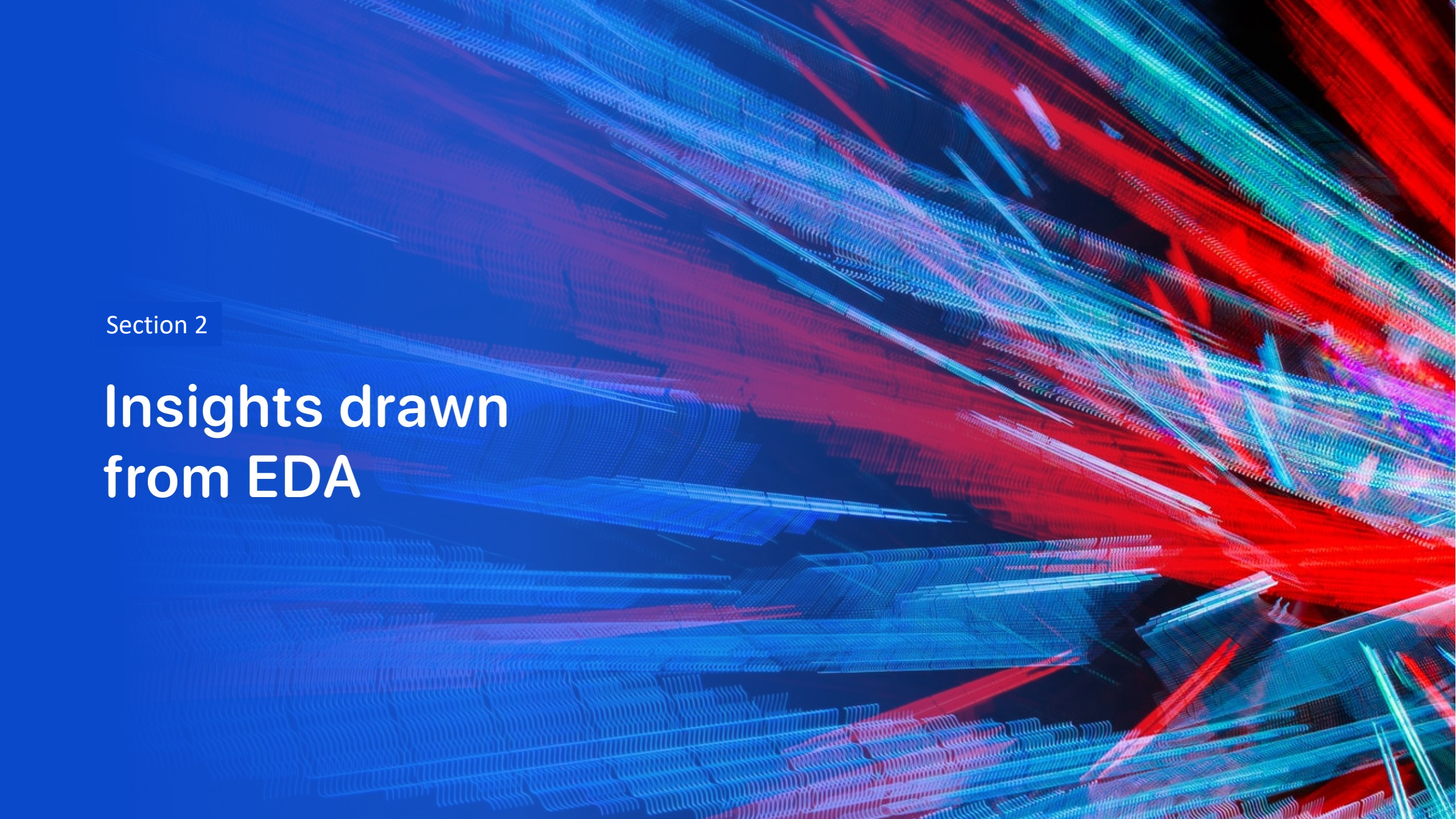
- Dashboard will include a **dropdown** list that lets users choose between all launch sites or a specific one.
- **A pie chart** will be displayed to show the percentage of successful and unsuccessful launches.
- Users can select the payload mass range using a **slider.**
- The **scatter chart** will display the correlation between payload mass and launch success by booster version.

Predictive Analysis (Classification)

1. Build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully.
2. Preprocess and standardize the data.
3. Split the data into training and testing data using train_test_split.
4. Train the model and perform Grid Search to find the best hyperparameters for the chosen algorithm.
5. Determine the model with the best accuracy using the training data.
6. Test different algorithms such as Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest neighbors.
7. Output the confusion matrix.

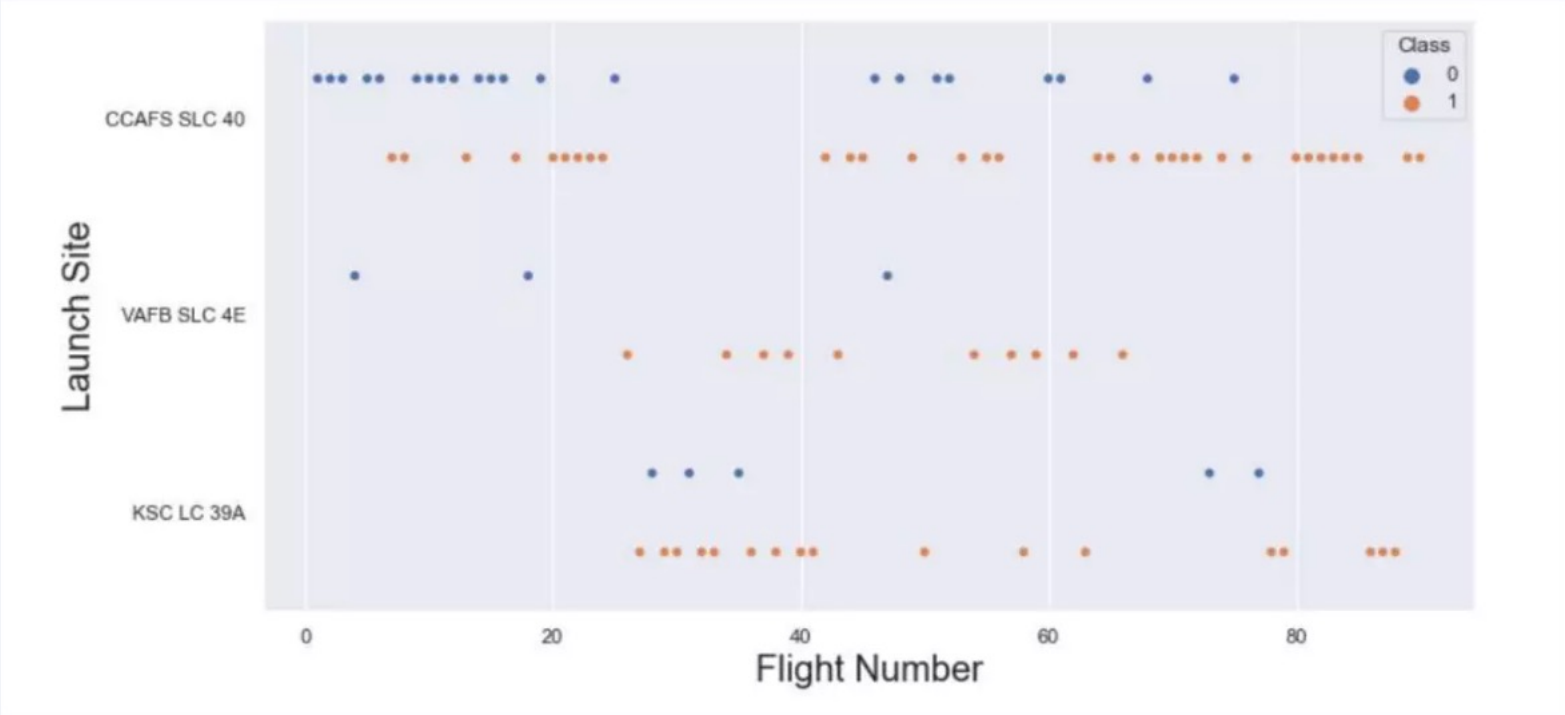
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

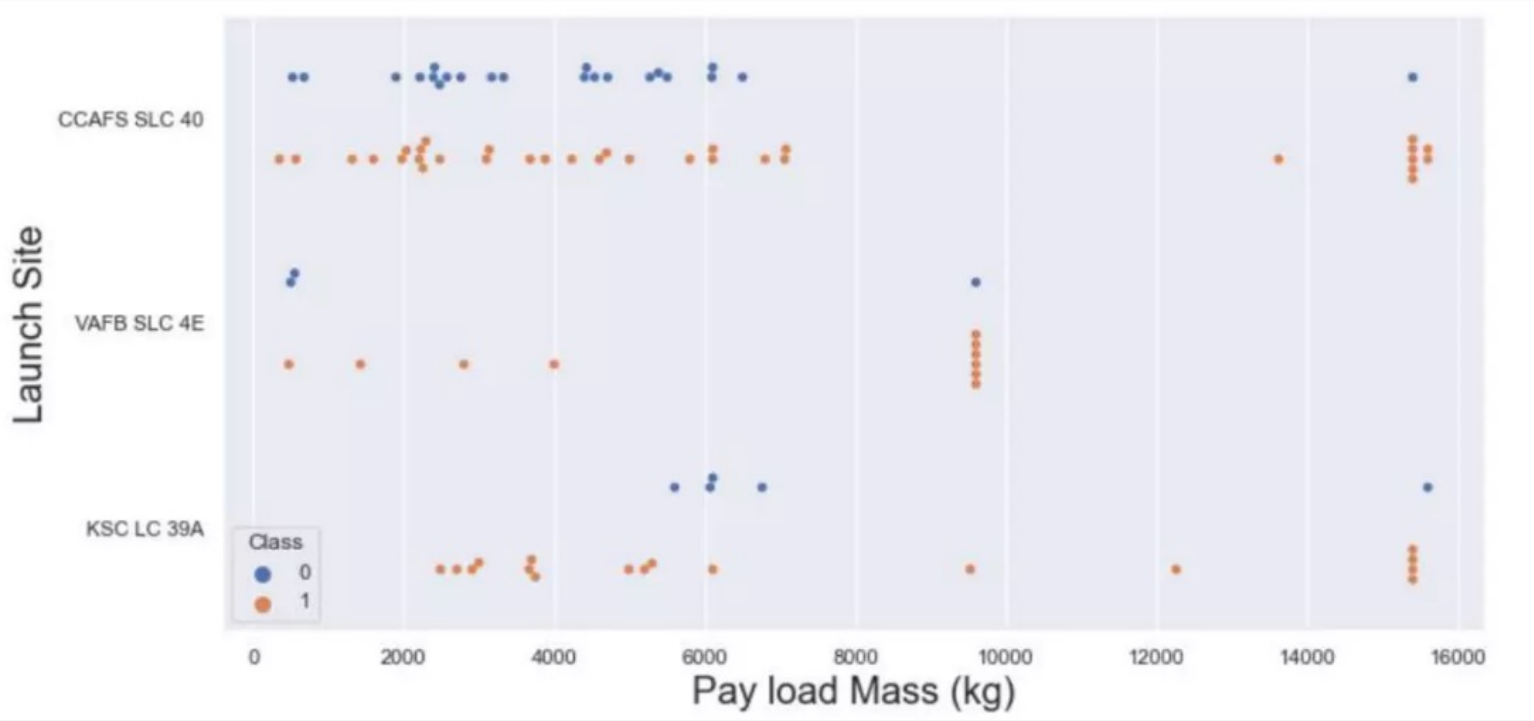
Flight Number vs. Launch Site

As the increase of flight number, which indicates a more recent launch, the success rate rises accordingly.



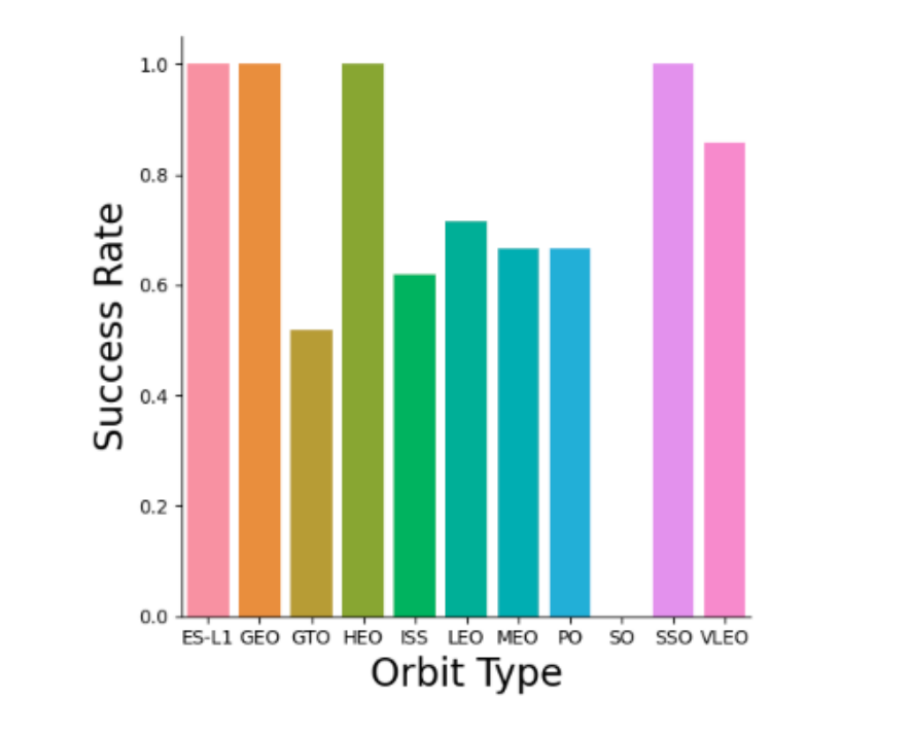
Payload vs. Launch Site

As the increase of payload, the success rate rises accordingly.



Success Rate vs. Orbit Type

As for the orbit type, SO performed worst and never succeeded, while ES-L1, GEO,HEO and SSO succeeded every time.

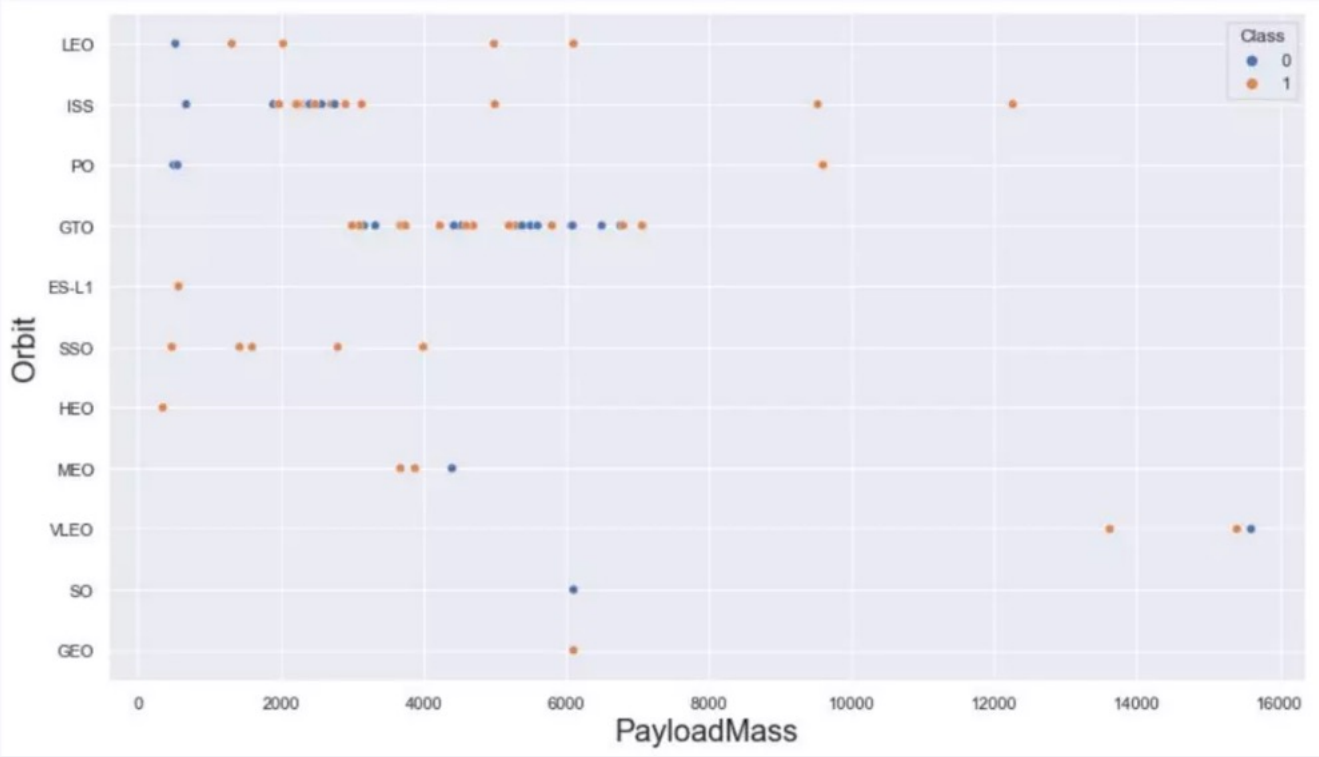


Flight Number vs. Orbit Type

As the increase of flight number, which indicates a more recent launch, the success rate rises accordingly in most of the orbit type.

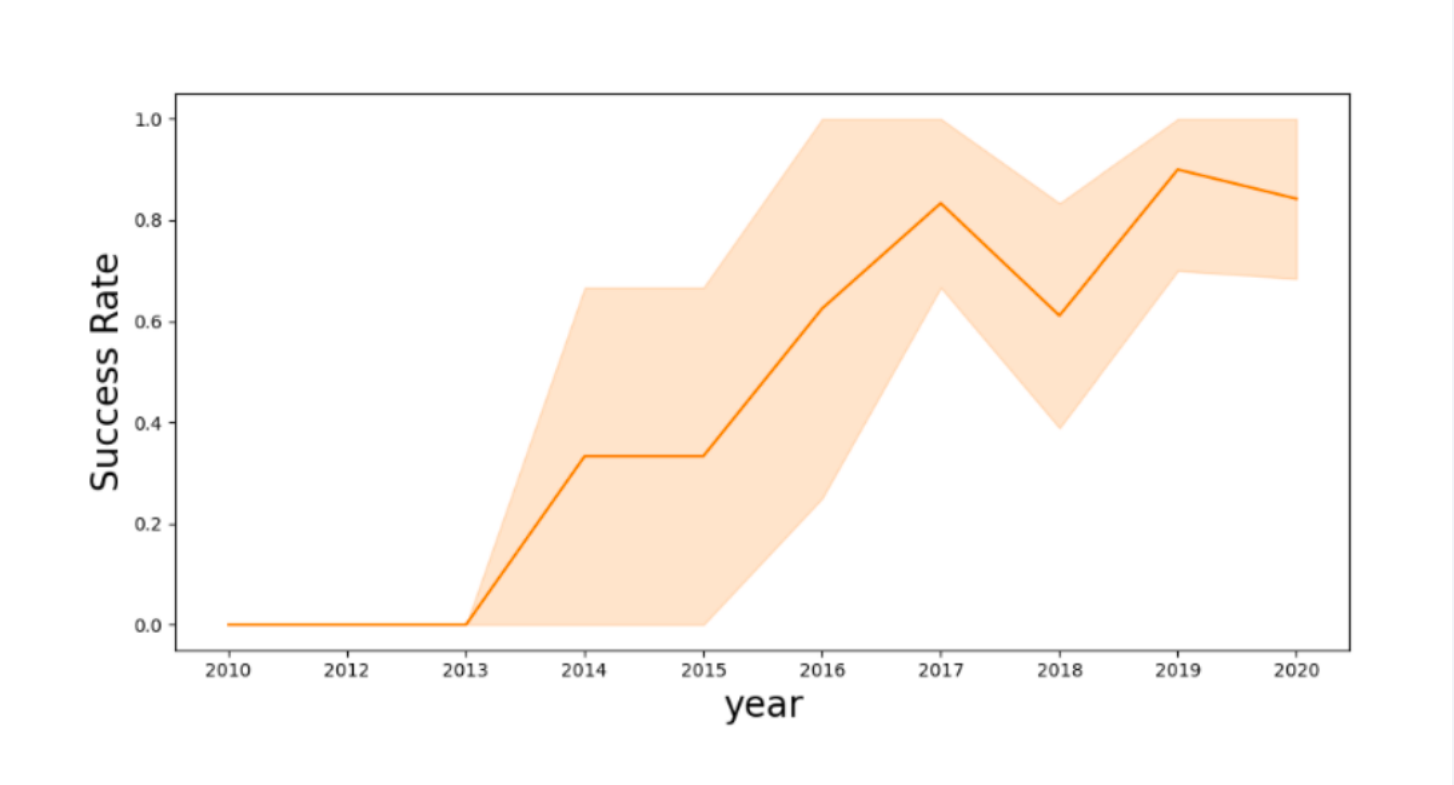
Payload vs. Orbit Type

As the increase of flight number, which indicates a more recent launch, the success rate rises accordingly in most of the orbit type.



Launch Success Yearly Trend

Basically, the launch success rate exhibited a rising trend, while in 2018, the success rate dropped.



All Launch Site Names

- CCAFS LC-40
- CCAFS SLC-40
- KSCLC-39A
- VAFB SLC-4E

All Launch Site Names

- CCAFS LC-40
- CCAFS SLC-40
- KSCLC-39A
- VAFB SLC-4E

Launch Site Names Begin with 'CCA'

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Apply SUM( ) statement in SQL
- 45,596 (total) carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

- Apply AVG( ) statement in SQL
- An average of 2,928 kg carried by boosters launched by F9 v1.1

First Successful Ground Landing Date

- Apply MIN( ) statement in SQL
- First successful ground landing occurs in 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Apply WHERE( ) statement in SQL
- Four ship landing: JSCAT-14,JSCAT-16,SES-10, SES-11 / EchoStar 105

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Apply COUNT( ) statement in SQL
- 1 Failure vs. 100 Success Mission

Boosters Carried Maximum Payload

- Apply MAX( ) statement in SQL
- Boosters List:
  - ✓ F9 B5 B1048.4
  - ✓ F9 B5 B1049.4
  - ✓ F9 B5 B1051.3
  - ✓ F9 B5 B1056.4
  - ✓ F9 B5 B1048.5
  - ✓ F9 B5 B1051.4
  - ✓ F9 B5 B1049.5
  - ✓ F9 B5 B1060.2
  - ✓ F9 B5 B1058.3
  - ✓ F9 B5 B1051.6
  - ✓ F9 B5 B1060.3
  - ✓ F9 B5 B1049.7

2015 Launch Records

- Apply month( ) statement in SQL

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

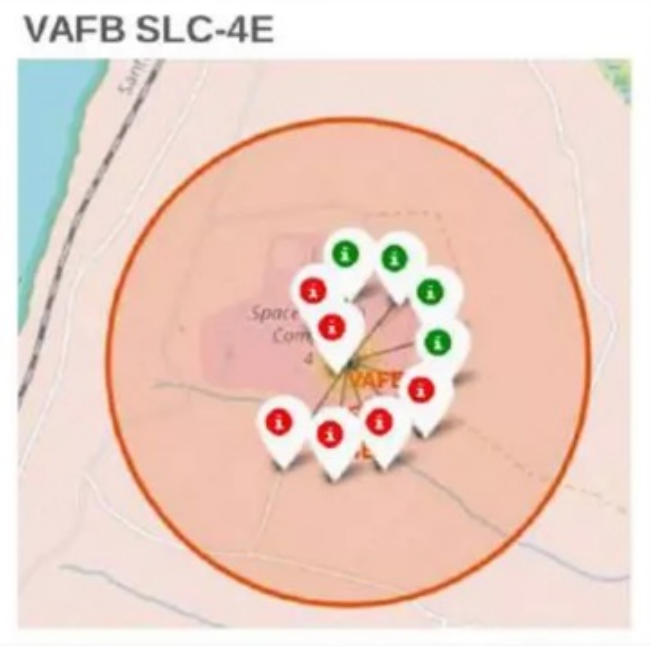
- Apply ORDER( ) statement in SQL

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

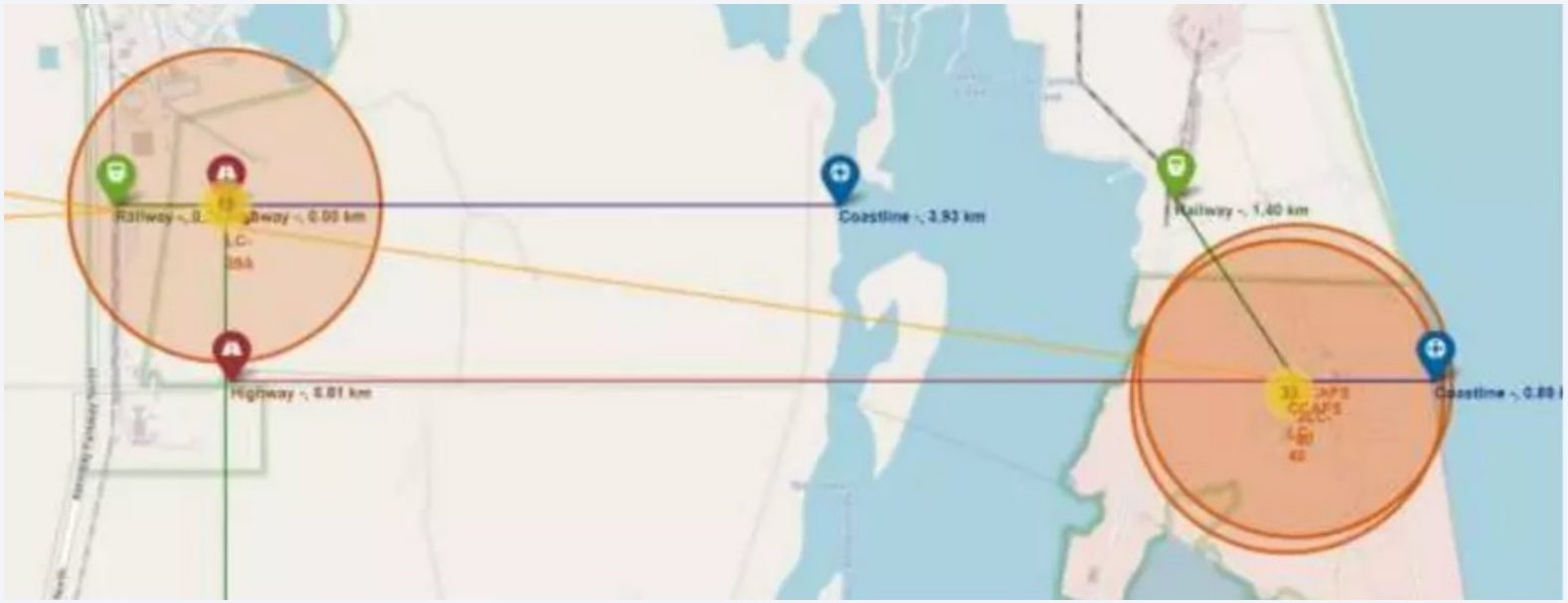

Section 3
Launch Sites
Proximities Analysis

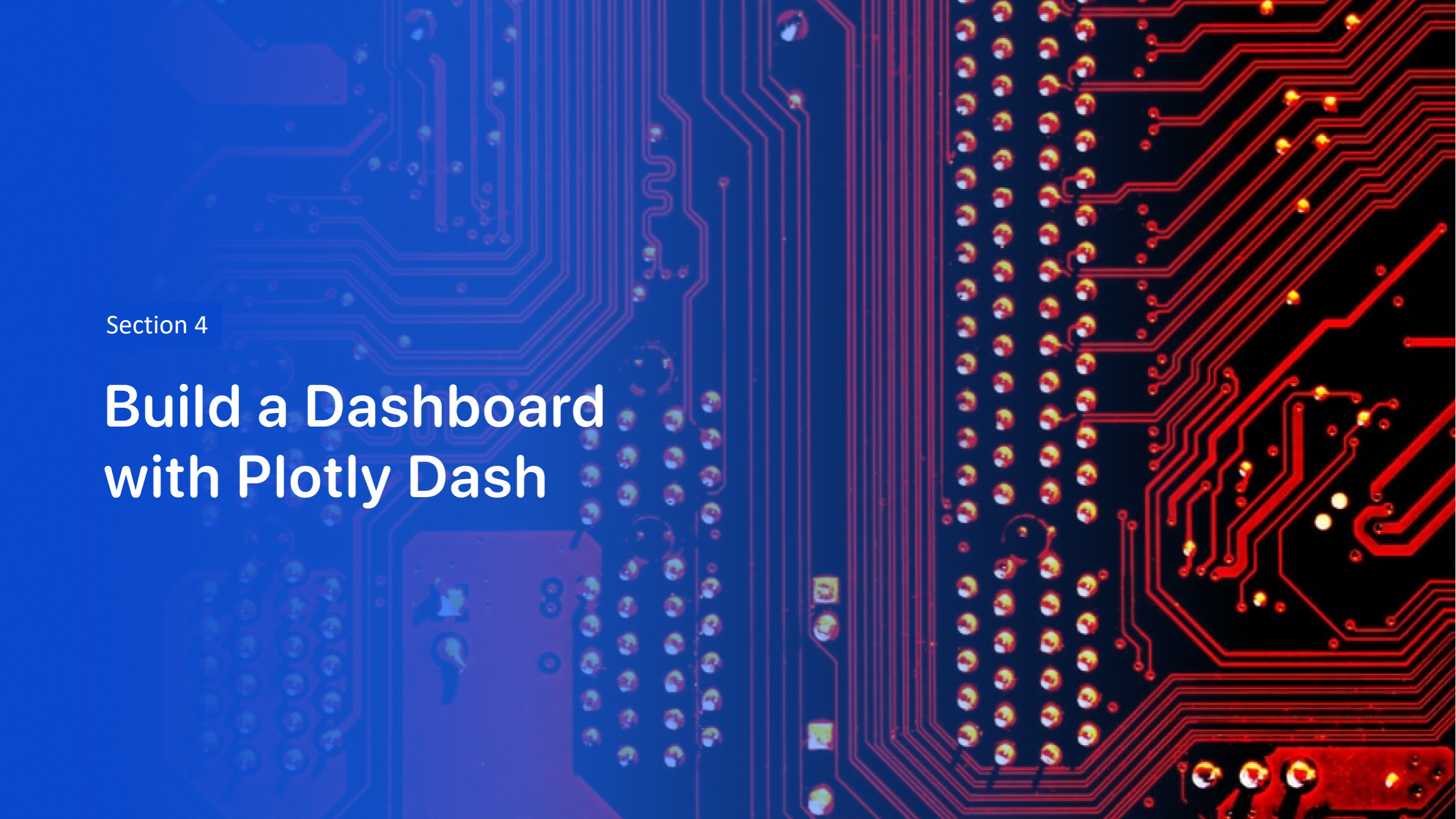
Launch Sites

Launch Outcomes



Distance




Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

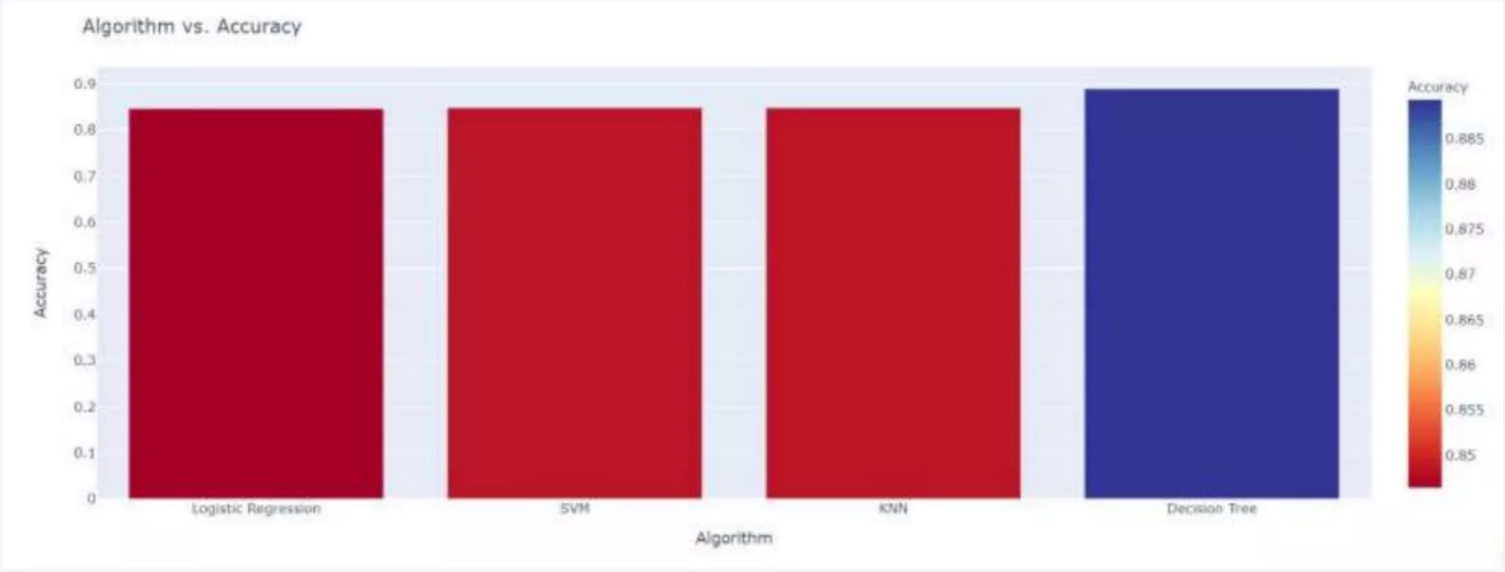
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

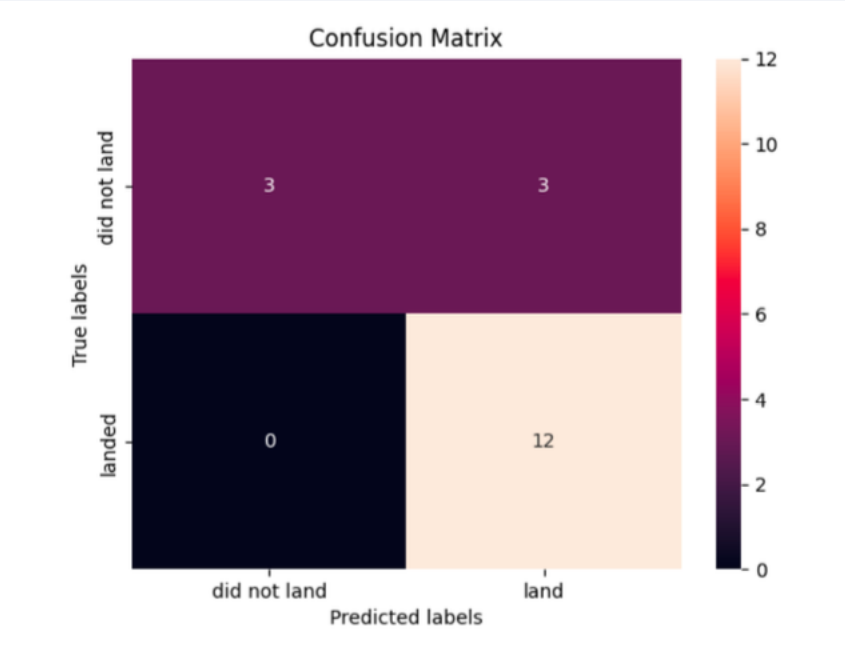
Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree performs best



Confusion Matrix



Conclusions

- Model Performance: The models showed similar performance on the test set, with the decision tree model slightly outperforming the others.
- Coast: All launch sites are situated close to the coast.
- Launch Success: The success rate of launches has increased over time.
- KSC LC-39A: This launch site has the highest success rate among all others and has a 100% success rate for launches with a payload of less than 5,500 kg.
- ES-L1, GEO, HEO, and SSO have a 100% success rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!